

Dynamic Modeling by Usage Data for Personalization Systems

Saeed R. Aghabozorgi¹, Teh Ying Wah²

Department of Information Science, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur – Malaysia

¹saeed@perdana.um.edu.my

²tehyw@um.edu.my

Abstract

With the extensive growth of data available on the Internet, personalization of this huge information becomes essential. Although, there are various techniques of personalization, in this paper we concentrate on using data mining algorithms to personalize web sites' usage data. This paper proposes an off-line model based web usage mining that is generated by clustering algorithm. Then, we will use users' transactions periodically to change the off-line model to a dynamic-model. This proposed approach will solve the problem of the decrease of accuracy in the off-line models over time resulted of new users joining or changes of behaviour for existing users in model-based approaches. Finally, we discuss the on-line model for user behaviour prediction in the web personalization system.

1. Introduction

Every day, millions of electronic pages, are added on hundreds of millions pages that are already on-line. With this substantial growth of available data on the Internet and because of its rapid and chaotic growth, the World Wide Web has emerged into a network of information without organizational structure. Furthermore, existence of abundant information in the network and the dynamic and heterogeneous nature of the web, web exploration has become a difficult process for most of the users. Causing users feel disoriented and sometimes lost in overloaded information that continues to expand. Besides, e-business and web marketing are rapidly developing and importance of anticipate the need of their customers is evident more than ever. Therefore, predicting the users' interests in order to improve the usability of web or so called personalization has become a necessity. According to Mobasher [6], "Web personalization can be described as any action that makes the web experience of a user personalized to the user's taste."

Usually, three types of data have to be managed in a web site: content, structure and log data. Content data

consist of whatever is in a web page; structure data refer to the organization of the content; usage data are the usage patterns of web sites. The application of the data mining techniques to these different data sets is at the basis of the three different research directions in the field of web mining: web content mining, web structure mining and web usage mining [1]. Srivastava and et al. [4] define the web usage mining as "Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications." Note that web usage mining differs from web structure mining and web content mining, in that web usage mining reflects the behaviour of humans as they interact with the Internet. Analysis of user behaviour in interaction with web site can provide insights leading to customization and personalization of a user's web experience. Because of this, web usage mining is of intense interest for e-marketing and e-commerce professionals [7].

Most of the existing works in the context of web usage mining try to classify a user while she is browsing the web site or using user's registration information [1]. Our main goal in this paper is based on this fact that in most web sites, it is not possible to perform an "on-line" modelling in term of quality (huge number of users or items). In addition, using a static model constructed in "off-line" mode may generate inaccurate result because the interests of users change over time and besides, new items and new users will be added to web site frequently. In this paper we present a model that is made in "off-line" mode, and then we change it to a dynamic model to predict user's interests in "online" mode. The novelty of our approach is that of using a 3 step recommendation system for personalization: The first stage is generation of an explicit model in "off-line" mode by data mining algorithms. This model is created by all user behaviours data collected during previous interactions with web site. The second stage is done periodically to avoid remodelling of generated model. Because the behaviour of users change over time and also, the total number of users of web site is increased, we use user's transactions after generating "off-line" model to adjust the model. This approach changes the static model to a

dynamic model. The last phase is carried out in "real-time" as a new visitor begins an interaction with the Web site. Data from the current user session is scored using the models generated offline, and recommendations generated based on this scoring. A general architecture for three step web mining is illustrated in figure 1.

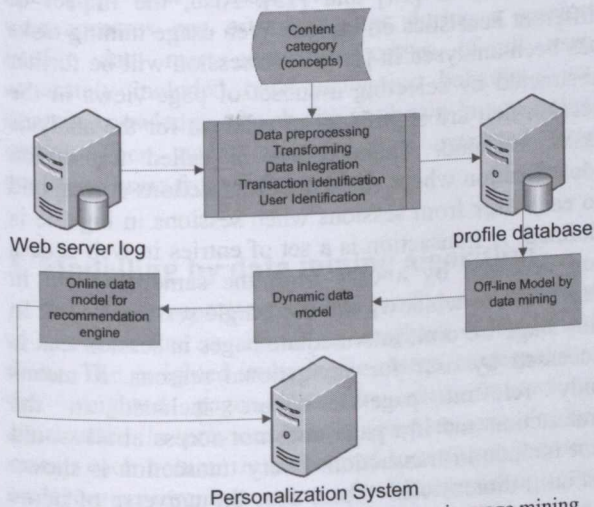


Fig. 1 Summary of the architecture and steps in usage mining

So far we have provided an overview of web mining and web usage mining for personalization. Then we showed the problem in standard techniques in personalization and a general overview of our approach in order to solve the problem.

The remaining of this paper is organized as follows: In section 2, we provide a detailed discussion of a host of data mining activities necessary for this process, including the pre-processing and integration of data from multiple sources, transaction identification as data source of web usage mining and a mathematical model of page-views in the process. In this Section we describe the data available for mining in the Web domain, specifically for the generation of user models. Then, we present the web usage mining concept and this access logs processing architecture in section 3. In this section the focus will be on Web usage mining where the goal is to use data collected from the user interactions with the Web in order to learn user models and to use these models for personalization in the future. We describe the clustering techniques used in generating the model of user behaviour in duration of user access to web site. Then, we define an approach for achieve an aggregate usage profile to present the usage of website. In section 4, we improve the model by changing the static model to a dynamic model. In the section 5 we show how a recommendation system use the model for combining the discovered knowledge with the current status of a user's activity in a Web site

to provide personalized content to a user. Finally, section 6 summarizes our conclusions and discussion on the current state and future direction of research in web usage mining.

2. Data Preparation

Variety types of data expressions could be used to model the co-occurrence of Web user behaviours, such as matrices, directed graphs and click sequences and so on. Different data expression models have different mathematical and theoretical backgrounds. In this paper, the user navigational behaviour is modelled by a usage matrix, in the form of transaction vectors.

2.1. Data Source

In Web mining, data can be collected at the server-side, client-side, proxy servers, or a consolidated web/business database. In [2], Srivastava et al. present a more detailed description of these data sources. To summarize, (i) Web server logs explicitly records browsing behaviour of site visitors, (ii) Client-side data collection can be implemented by using a remote agent or by modifying the source code of an existing browser (iii) and Web proxies act as an intermediate level of caching between client browsers and Web servers.

The primary data sources used in Web usage mining are user queries, registration data and the server log files which include Web server access logs and application server logs [3,8]. In this paper, Web server log file and domain knowledge of websites is used as data source. Web server access logs store data pertaining to access of the website. Each hit against the server, corresponding to an HTTP request, generates a single entry and stores it in web server access log.

2.2. Pre-processing

The data collected from different data sources are stored as a data mart in a data warehouse (such as those described in [9]). The information provided by these data sources in data warehouse is used to construct several data abstractions, namely users, page-views, click-streams, and server sessions [3]. Analyses of these data will lead to extract model of user behaviour in interaction with the web site. Hence, before processing of the data we need to perform some pre-processing tasks. Pre-processing consists of converting usage data contained in the various available data sources into the data abstractions necessary for pattern discovery. In this section we describe pre-processing steps performed on web log files including user identification, page-view identification, transaction identification, and data storage in a repository.

Meanwhile, we use domain knowledge of web site or concept of web site for abstraction and cleaning input file. Each web-site, depending on its application, provides information about one or more concepts. For example, amazon.com includes concepts such as Books, Music, Video, etc. The web pages within a website are categorized based on the concept(s) to which they belong. A concept space, or simply concept, in a website is defined as the set of web pages that contain information about a certain concept [10].

2.3. User Identification

It is important to notice that the data collected by server logs may not be entirely reliable because some page views may be cached by the user's browser or by a proxy server [14]. The data recorded in server logs reflects (possibly concurrent) the access of a Web site by multiple users, and only the IP address, agent and server side click-stream are available to identify users and server sessions. In a Web server log, all requests from a proxy server have the same identifier, even though the requests potentially represent more than one user. On the other hand, the Web server can also store other kinds of usage information such as cookies, which are markers generated by the Web server for individual client browsers to automatically track the site visitors. But, a user may access the Web through different machines, or use more than one browser at one time. In practice, it is very difficult to uniquely and repeatedly identify users. In this paper we identify each user through cookies, logins, or IP/agent analysis. We assume the existence of a set of m users, $U = \{u_1, u_2, \dots, u_m\}$ where each user is defined as a single individual that is accessing files in Web servers (page-views) through a browser.

2.4. Page-view Identification

Page-view is an aggregate representation of a collection of Web objects contributing to the display on a user's browser resulting from a single user action (such as a click-through) [11]. We consider each item in data mining as a page-view in web usage mining that in the simplest case, each HTML file has a one-to-one correlation with an item. If each page-view is shown as p_i then we define universe of all pages in a web site as $P = \{p_1, p_2 \dots p_n\}$.

2.5. Transaction Identification

After definition of each user, the click-stream for each user must be divided into sessions. A server-session (or visit) is the click-stream (sequence of page-views) by a single user for a particular purpose in a

single visit. The end of a server-session is defined as the point when the user's browsing session at that site has ended [11]. But because it is not clear when the user has left the Web site, a timeout is often used as the default method of breaking a user's click-stream into sessions. Various heuristics for defining sessions especially with anonymous users have been identified and studied in [14] and [15]. Also, the impact of different heuristics on various Web usage mining tasks has been analysed in [16]. Each session will be further abstracted by selecting a subset of page-views in the session that are significant or relevant for the analysis tasks at hand. This process is called transaction identification where we extract transactions correspond to each user from sessions when sessions in log file is defined. A transaction is a set of entries in web access log accessed by a user from the same machine in defined time windows and in a single session [sa10]. In this stage we omit intermediate pages in session that is accessed by user for navigational reasons. It means only relevant page-views are included in the transaction and if a page does not access at all would not include in transaction. Every transaction is shown as an n -dimensional vector over the universe of items or especially in web mining as page-view vectors, i.e. Transaction t_i is shown as

$$t_i = \langle w(p_1, t_i), w(p_2, t_i), \dots, w(p_n, t_i) \rangle \quad (1)$$

Which is a group of Web page accesses $w(p_i, t_i)$ where w represents weight of p_i in transaction t_i . The weight is determined by features of page-view such as page duration, purchase a product, rate and etc. In this paper we define transactions for each user to apply clustering algorithms and dynamically create meaningful clusters of users, based on underlying model of the user's browsing behaviour.

2.6. Storing User Profiles in Repository

In this stage we will aggregate all transactions pertaining to a user during prior interactions with web sites to generate profile of each user. The profile of each user can be short-term profile based on only one visit of user from site or aggregation of all transactions through the all visits of user from website that is called long-term profile. Mobasher and et al [11] represent this profiles as an $m \times n$ matrix. They show the profile for a user $u \in U$ as an n -dimensional vector of ordered pairs:

$$u^{(n)} = \langle (i_1, s_u(i_1)), (i_2, s_u(i_2)), \dots, (i_n, s_u(i_n)) \rangle \quad (2)$$

where i_j 's $\in I$ and s_u is a function for user u assigning (possibly null) interest scores to items. In this stage an agent is defined that for each valid item (page-view) assigns corresponding concept (category) based on site

structure information present on the page's URL or based on the defined structure in content management systems or based on keywords included in page-views. If in our research we consider the significant pages for different page's type, the agent performs the definition of more significant pages than others as well. In this stage user profiles are stored in a data mart.

Each page-view is associated to one category and a user accesses one or more items during a session. Profile data mart contains accesses of all users accurately included both navigation behaviour and domain knowledge of web site as shown in conceptual model. Then, useful knowledge is extracted from profile of users from the data mart.

3. Modelling by data mining algorithm

The data in repository of user profiles is considered as a matrix that is shown as weighted collection of items. The weighted items in the matrix correspond to the weighted pages-views in user transactions constructed in previous phase (preparation). In this matrix each column is associated to a page-view and each row represents a user profile. The weights in the matrix are determined in a number of ways, for example, binary weights can be used to represent existence or non-existence of a product-purchase or a document access in the user transactions. Additionally, the weights can be a function of the duration of the associated page-view in order to capture the user's interest in a content page [12]. The weights may also, in part, be based on domain-specific significance weights (for example navigational pages may be weighted less heavily than content or product-oriented page-views). With this representation, the user profile is considered as a page-view vector that each item in the vector has an associated weight that represents its significance in the profile.

At this stage, an explicit "off-line" model is generated by data mining algorithms. This model is created by user behaviour data collected during previous interactions with web site. A number of data mining algorithms can be used for "off-line" model building such as Clustering, Classification, Association Rule Discovery, Sequential pattern Discovery, Markov models, hidden (latent) variable models. Each of these algorithms generates a model that is used for analysing the access pattern of users. We use a clustering algorithm to partition users based on their profiles. Clustering is an undirected or unsupervised method, meaning that the analyst need not define a target variable. Instead, the data mining algorithm searches for patterns and structure among the variables [7]. After applying clustering algorithm, each cluster represents a group of users with similar navigation

patterns. There are various algorithms which are used for clustering user profiles. For our approach we use traditional k-mean algorithm for cluster our users' profiles. Standard clustering algorithms, such as k-means, generally partition this space into groups of transactions that are close to each other based on a measure of distance or similarity [12]. K-mean algorithm is an algorithm to cluster N objects (data point) based on attributes into K partitions, that $k < N$. It assumes that the object attributes form a vector space. The objective that it tries to achieve is to minimize total intra-cluster variance, or, the squared error function [18]:

$$V = \sum_{i=1}^k \sum_{U_n \in C_i} (U_n - \mu_i)^2 \quad (3)$$

Where $U_n \in C_i$ is a vector representing the n th data point and in our context the user profile and there are k clusters C_i , $i = 1, 2, \dots, k$, and μ_i is the centroid (geometric centroid of data points) or mean point of all the points $U_n \in C_i$.

Implementation of this algorithm is included below stages: 1) Selecting k point randomly as k centroid of clusters 2) For each data point x , compute its membership in clusters by choosing the nearest centroid 3) For each centroid, recomputed its location based on members

Clustering of user's profiles table is based on having something in common in user transactions. With applying K-mean algorithm on user profiles a set of clusters is achieved that each cluster C_i includes a subset of the users with similar navigational patterns (previous behavior in interaction with site) or on the other word same transactions. In process of clustering, users' profile are clustered based on the similarity of the interest scores for page-views across all users, or based on similarity of their content features. Each cluster is included hundred of user profiles. On the other hand, each user profile is involving hundreds of page-view references and it should be reduced to represent clusters into weighted collections of page-views. At first, we try to extract separate clusters with high confidence, because later we will use the aggregated of these clusters as centroid of clusters in dynamic model. Then, we use profile aggregation as a method for derivation of aggregated profiles from clusters. As mentioned, because each cluster includes transactions involving many page-views, an aggregated profile is generated from each cluster. To aggregate transactions we use concept of PACT approach [12] (Profile Aggregations based on Clustering Transactions). This approach is a technique analogous to concept indexing methods used to extract document cluster summaries in information retrieval and filtering

[13]. But here we use aggregate of profiles instead of all transactions in a cluster to generate aggregated profiles. The aggregate profiles are generated based on centroid of each cluster. For each cluster $C_i \in T_C$, the centroid (m_c) is computed by finding the ratio of the sum of the page-view weights across profiles in C_i to the total number of users in the cluster.

$$weight(p, pr_c) = \frac{\sum_{U \in C} w(p, t)}{N} \quad (4)$$

4. Dynamic data model

Generating the user profile model from transactions is the off-line part of modelling. Because new users and new items increase and in addition, the current users' interest change over time and subsequently users will do new transactions with the web site, we use an approach to improve clusters by adjusting model through new transactions. In this approach we consider the behaviour of user since making offline model and calculate the probability of assigning a user to other clusters. To this aim, at first, interesting of each cluster about each category in a predefined interval is calculated by counting the number of request sent to each category from each cluster. Then we will calculate the probability of interesting of each user about each category by counting the number of request from a user regarding the probability of cluster's interesting. Finally, we assign a user to a cluster that has maximum probability.

The stage of attributing categories to clusters can be done by considering the initial clusters achieved through the k-mean algorithm and by number of transactions done by the users of each cluster with a category domain after generating static model in offline mode (time interval). Each category are related to the cluster with a probability value equals the ratio of the number of requests sent to the category from users of the cluster over all clusters. Although, this type of making correlation is not completely accurate, but if the period of time be long and the number of transactions be big enough, it can be reliable. Furthermore, there are some techniques for finding clusters that consider the content categories of pages. To shed light this issue, the frequent item set is an example of finding relation among user clusters and content categories [17].

For calculating relation between clusters and category P_j in ontology domain, U_i is defined as universe of registered users and C_i as initial cluster produced by a clustering algorithm. Given the total number of users belonging to each cluster after the clustering N_i , total number of requests of U_k in cluster C_i for each category RU_{kj} , the requests of cluster C_i for

each category RC_{ij} is total number of requests from the cluster divided by total number of users within the cluster: (it is normalized requests sent for each category over total number of users within each cluster)

$$RC_{ij} = \frac{\sum_{U_k \in C_i} RU_{kj}}{N_i} \quad (5)$$

A matrix of cluster-category represents correlation among clusters and categories. Each cell represents probability PC_{ij} that is relation between category P_j and cluster C_i is computed:

$$PC_{ij} = \frac{RC_{ij}}{\sum_{i=0}^n RC_{ij}} \quad (6)$$

Then, we find the change of users' interests by considering transactions made after clustering. If the requests that user has sent through transactions for content categories is different with the categories that his cluster belongs, the cluster of user will change. In the other word, the probability that a user belongs to a cluster is computed by counting the number of times each user sends request for every category after constructing the offline model.

$$PUC_{ki} = \frac{\sum_{U_k \in C_i} RU_{kj} * PC_{ij}}{N_i} \quad (7)$$

The last action is assigning each user to the cluster that user has maximum probability that is belong to that. Therefore, if the majority of a user's request is for a category that has not a high correlation with her cluster, user will drop in another cluster. The model that is constructed, considers the changes in the behavior of users after off-line modeling and do not need to apply clustering algorithm on all data. Therefore, this model can be used as a dynamic model without scalability shortcomings in traditional on-line models.

5. Prediction

In order to address the requirement of effective web navigation, web sites provide personalized recommendations to the end users. The main purpose of personalization is prediction a set for the active user. This set must be including the objects (links, advertisements, text, products, etc) that are same as active user profile. This part of personalization must be done in online mode. For this mean, the system needs the history of user and transactions with web site or at least the track of current visit.

For recommendation systems, different prediction algorithms are used. For the registered users we compare the centroid of the cluster that user belongs with the matrix of user-item generated to finding interest items in repository to generate recommendation. The items in the centroid vector that have higher weight rather than the interest items in the matrix are interest items. For recommend to anonymous users, we will use nearest neighbour to predict items for users. To this aim, a matrix of user-item is generated to finding interest items for a user. The columns of this matrix is items of centroid vector that is most similarity with the user-session and the value of cells is the items that user has shown that interested in short session or long sessions. The interest item for the user is the null value in matrix.

If the data collection procedures in the system include the capability to track users across visits, then the recommendation set can represent a longer term view of potentially useful links based on the user's activity history within the site. If, on the other hand, profiles are derived from anonymous user sessions contained in log files, then the recommendations provide a "short-term" view of user's navigational history [12]. These recommended objects are added to the last page in the active session accessed by the user before that page is sent to user's browser.

6. Conclusions

Generating an offline model using usage data as an information source will dramatically improve the recommendation performance and the online response efficiency in personalization systems. On the other hand, because the behaviour of users changes over time, the model must regenerated using all usage data that in term of quality is hard. In this paper a new approach for personalization system by web usage mining was presented. We applied a data mining technique for generating an offline model of user's access pattern in visit websites. Then we made an aggregated user profile that can be used as main component of usage-based recommender system for prediction and recommendation for personalization. The advantage of our approach is generating a model in off-line mode and then changes the model to a dynamic model. Because the time consuming part of modelling is done by offline mode, and the model is adjusted periodically, personalization system won't need regenerating of model over time. Although this approach consider new users and changing in interest of users and solve this problem to a large extend, in some parts it needs more works. For example considering user's demographic in clustering and

finding the best number of clusters in k-mean clustering for our approach.

7. References

- [1] F. Zhang and H. Chang. "Research and development in web usage mining system- key issues and proposed solutions: a survey". In *First IEEE Int. Conf. on Machine Learning and Cybernetics Proceedings*, pages 986-990, Nov. 2002.
- [2] J. Srivastava, R. Cooley, M. Deshpande, P.-N. Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations, 1(2), Jan 2000.
- [3] WWW Committee Web Usage Characterization Activity, <http://www.w3.org/WCA>, Web Characterization Terminology & Definitions Sheet, W3C Working Draft, May 1999.
- [4] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan, Web usage mining: discovery and applications of usage patterns from web data, SIGKDD Explor., 1(2), Jan. 2000.
- [5] R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining World Wide Web browsing patterns. *Journal of Knowledge and Information Systems*, (1)1, 1999.
- [6] B. Mobasher, R. Cooley, and J. Srivastava. Automatic personalization based on web usage mining. *Communications of the ACM*, 43(8):142-151, 2000.
- [7] Z. Markov, and D. T. Larose, *Data mining the Web : uncovering patterns in Web content, structure, and usage*, Hoboken, N.J.: Wiley-Interscience/John Wiley & Sons, 2007.
- [8] M. Baglion1, U. Ferrara2, A. Romei1, S. Ruggieri1, and F. Turini1, Preprocessing and Mining Web Log Data for Web Personalization, 2003
- [9] M. Sweiger, M.R. Madsen, J. Langston, and H. Lombard. Clickstream Data Warehousing. John Wiley & Sons, 2002.
- [10] C. Shahabi, and F. Banaei-Kashani, "A Framework for Efficient and Anonymous Web Usage Mining Based on Client-Side Tracking," *WEBKDD 2001 - Mining Web Log Data Across All Customers Touch Points*, pp. 141-155, 2002.
- [11] P. Brusilovsky, A. Kobsa, and W. Nejdl, *The adaptive web : methods and strategies of web personalization*, Berlin ; New York: Springer, 2007.
- [12] B. Mobasher, H. Dai, T. Luo *et al.*, "Discovery and evaluation of aggregate usage profiles for Web personalization," *Data Mining & Knowledge Discovery*, vol. 6, no. 1, pp. 61-82, 2002.
- [13] G. Karypis, E-H. Han. Concept indexing: a fast dimensionality reduction algorithm with applications to document retrieval and categorization. Technical Report #00-016, Department of Computer Science and Engineering, University of Minnesota, March 2000.
- [14] Cooley, R., Mobasher, B., Srivastava, J.: Data preparation for mining World Wide Web browsing patterns. *Journal of Knowledge and Information Systems* 1(1) (1999) 5-32
- [15] Spiliopoulou, M., Mobasher, B., Berendt, B., Nakagawa, M.: A framework for the evaluation of session reconstruction heuristics in web usage analysis. *INFORMS Journal of Computing - Special Issue on Mining Web-Based Data for E-Business Applications* 15(2) (2003)
- [16] Zaiane, O.R., Srivastava, J., Spiliopoulou, M., Masand, B., eds.: *Proceedings of WEBKDD 2002 - Mining Web Data for Discovering Usage Patterns and Profiles*. Volume 2703 of LNCS. Springer Berlin / Heidelberg (2003) 159-179
- [17] P. Batista, and M. J. Silva, "Mining Web Access Logs of an On-line Newspaper."
- [18] J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations", *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1:281-297